IEEE Seminar on Digital Signal Proc. for Communication
Herzelia, Israel 1987

A Unified Framework for Residual Coding of Speech

E. Ofer A. Dembo D. Malah

Electrical Engineering Department
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

Abstract

A unified mathematical framework for several residual speech coders proposed in the literature is presented. In the unified approach the excitation signal to the synthesis filter is represented by a linear combination of a small number of vectors taken from a given set. The vector set is known both at the transmitter and at the receiver. The transmitted parameters are the indices of the chosen vectors from the set and the coefficients needed to linearly recombine the vectors. It is shown that several residual coding schemes such as Multipulse, CELP, and MRM, as well as coders which transform the residual signal, all fall into this unified framework.

1. Introduction

Residual speech coders are typically designed to operate at transmission rates of 4.8 - 16 Kbps. These coders have the feature of coding both the LPC residual, which is used as the excitation to the synthesis filter, and the LPC parameters which make up the synthesis filter [1-3].

In recent years, some new residual coders were proposed. In the Multipulse scheme [4], the residual signal is represented by a small number of pulses. The pulses locations and amplitudes are coded and transmitted in addition to the LPC parameters. Following the multipulse scheme other schemes have appeared which attempt to represent the residual in a simpler manner. Some examples are: Maximum Residual Magnitude (MRM) [5], Regular Excitation [6], Thinned-Out Residual [7], etc.

Another recently proposed residual coder is the CELP (Code Excited Linear Prediction) coder [8] in which each segment of the residual signal is represented by an appropriate block of white Gaussian noise selected from a given dictionary (code-book). In other recent residual speech coders, the speech residual is first transformed and then coded [9],[10].

In this work, we present a unified mathematical framework for some of the above mentioned coders. In this scheme, which is based on a vector-expansion of the LPC Residual, the residual signal is represented by a linear combination of a small number of vectors taken from a given vector set.

In section 2, a mathematical presentation of the vector-expansion idea is given. Section 3 demonstrates how some known residual speech coders are particular examples of the unified framework. Section 4 then presents the derivation of an optimal vector set for the proposed vector-expansion scheme, and Section 5 presents simplified residual-transform schemes which emerge from the optimal scheme in Section 4.

2. Vector Expansion of the LPC Residual

In the proposed vector-expansion scheme, the excitation vector to the synthesis filter is represented by a linear combination of a small number of vectors, taken from a given vector set. The vector set is known at both the transmitter and the receiver. The transmitted parameters are the indices of the selected vectors from the set and the coefficient needed to linearly recombine the vectors.

Let V be a set of M vectors of length N, each. Given the original speech vector $\underline{s}$ - constructed from consecutive N samples of the input speech, and the LPC synthesis filter coefficients $\underline{a}$. We represent the excitation $\underline{u}$ as a linear combination of a small number of vectors from V, i.e.

$$\underline{u} = \sum_{i=1}^{k} x_i \underline{v}_i, \quad \underline{v}_i \in V, \quad k \ll M \tag{1}$$

The vectors $\underline{v}_i$ are to be selected such that the weighted mean square error (WMSE) between the original and the reconstructed speech signals is minimal. The WMSE is given by:

$$E_w = \sum_n [(s_n - \hat{s}_n) * w_n]^2 \tag{2}$$

where s_n is the original speech signal, $\hat{s}_n$ is the reconstructed speech signal, w_n is the weighting filter impulse response, and the summation is assumed here to be over the N samples of the frame. The transfer function of the weighting filter is given by [2,4]:

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 + \sum_{k=1}^{p} a_k z^{-k}}{1 + \sum_{k=1}^{p} a_k \gamma^k z^{-k}} \tag{3}$$

where the constant γ, $0 \le \gamma \le 1$, controls the amount of deemphasis of the error spectrum in the formant regions and is matched to the frequency masking properties of the human ear [2]. Representing the reconstructed signal as:

$$\hat{s}_n = u_n * h_n + l_n \tag{4}$$

where u_n is the excitation, h_n is the synthesis filter impulse response, and l_n is a "leftover" signal generated by the filter memory from previous frames (and needs not to be approximated, as it is also known at the receiver). Thus, from (2):

$$E_w = \sum_n [(e_n - u_n) * \tilde{h}_n]^2 \tag{5}$$

where e_n is the residual signal resulting from passing $(s_n - l_n)$ through the inverse filter $A(z)$ with zeroed-out memory, and $\tilde{h}_n$ is the impulse response of the weighted synthesis filter which is $h_n * w_n$.

Defining R to be the $N \times N$ matrix with elements r_{ij} given by:

$$r_{ij} = \sum_n \tilde{h}_{n-i} \tilde{h}_{n-j} \qquad 0 \le i \le N-1, \quad 0 \le j \le N-1 \tag{6}$$

results in

$$E_w = (\underline{e} - \underline{u})^T R (\underline{e} - \underline{u}) \tag{7}$$

where $\underline{e}$ is the residual vector and $\underline{u}$ is the excitation vector.
Substituting $\underline{u}$ by its vector expansion (1) gives:

$$E_w = (\underline{e} - \sum_{i=1}^{k} x_i \underline{v}_i)^T R (\underline{e} - \sum_{i=1}^{k} x_i \underline{v}_i) \tag{8}$$

Now, let Q be an $N \times k$ matrix, having as its columns the selected k vectors from V. For a given matrix Q, we obtain the following expression for the WMSE:

$$E_w^Q = (\underline{e} - Q\underline{x})^T R (\underline{e} - Q\underline{x}) \tag{9}$$

where the elements of the vector $\underline{x}$ are the linear combination coefficients x_i, $i = 1, 2, \ldots, k$.
Solving for $\underline{x}$ which minimizes E_w^Q, results in:

$$\underline{x}_{opt} = (Q^T R Q)^{-1} Q^T R \underline{e} \tag{10}$$

Substituting (10) into (9) gives the expression for the minimal error for the given matrix Q:

$$E_{w_{\min}}^Q = \underline{e}^T R \underline{e} - (\underline{e}^T R Q)(Q^T R Q)^{-1} (Q^T R \underline{e}) \tag{11}$$

For each speech frame, we can compute $\underline{e}$ and R and then have to find $Q \in V$ such that ΔE_w^Q will be maximal, where:

$$\Delta E_w^Q = (\underline{e}^T R Q)(Q^T R Q)^{-1} (Q^T R \underline{e}) \tag{12}$$

3. Particular Examples

The residual expansion scheme presented in the previous section provides a unified framework for some residual speech coders which were recently proposed. These coders, which we will review here briefly, differ from each other by the chosen vector set V and the way the expansion basis vectors are selected from it.

3.1 Multipulse LPC

Let V be chosen to be equal to the $N \times N$ unity matrix, i.e.,

$$V = I_N \tag{13}$$

In this case the excitation signal is a linear combination of k unity vectors. For this choice of V, eq. (12) does not reduce to a form which allows us to select all the k vectors simultaneously. Thus, an iterative suboptimal solution is typically used. This choice of V, and the use of an iterative algorithm, is equivalent to representing the excitation signal with only k pulses positioned in the frame in positions and with amplitudes which are calculated iteratively - one pulse at a time. This scheme is known as Multipulse LPC and was first suggested by Atal & Remde [4]. Using (12) shows that in the j-th iteration of the algorithm we should choose $Q = \underline{v}_i$, $v_i \in V$ such that:

$$\Delta E_w^j = (\underline{e}^{jT} R \underline{v}_i)^2 / r_{ii} \tag{14}$$

is maximized (over all vectors in V). r_{ii} is the i-th diagonal element of R, and $\underline{e}^j$ is the residual vector $\underline{e}$ updated to take into account all the chosen vectors up to this point. Thus,

$$\underline{e}^{j+1} = \underline{e}^{j} - x_i \underline{v}_i \; ; \qquad \underline{e}^{o} = \underline{e} \tag{15}$$

Implementing (14) is simple since multiplying the matrix R by $\underline{v}_i$ means choosing its i-th column.

3.2 CELP

In the CELP scheme [8], a large vector set is used (a dictionary). The vectors here are usually blocks of white Gaussian noise and only one vector is chosen per frame. Using (12) shows that for each frame a single vector $\underline{v}$, $\underline{v} \in V$ should be chosen such that

$$\Delta E_w^{\underline{v}} = (\underline{e}^T R \underline{v})^2 / (\underline{v}^T R \underline{v}) \tag{16}$$

will be maximized over all the possible vectors in V. Thus, the process of selecting a vector in Multipulse and CELP is quite similar.

3.3 Generalized MRM (GMRM)

Examination of (12) shows that if we use a vector set V having the property that for any Q chosen from it the matrix $(Q^T R Q)$ is diagonal, then we can select from V all the vectors in Q simultaneously.

Let R be calculated by the 'covariance' type method. R can then be represented as:

$$R = \tilde{H}^T \tilde{H} \tag{17}$$

where $\tilde{H}$ is the $N \times N$ lower triangular matrix, with an (i, j) element $\tilde{h}_{i-j}$, for $i \geq j$. Let V chosen to be:

$$V = \tilde{H}^{-1} \tag{18}$$

Then, since Q is a column-submatrix of V we obtain $Q^T R Q = I_k$, and hence, from (12):

$$\Delta E_w^{Q} = (\underline{e}^T R Q)(Q^T R \underline{e}) \tag{19}$$

This expression will be maximized by choosing the vectors composing Q according to the indices in V which correspond to the k largest elements in absolute value from:

$$\underline{y} = V^T R \underline{e} = \tilde{H} \underline{e} \tag{20}$$

Hence, y can simply be found by passing the residual vector $\underline{e}$ through the weighted synthesis filter $H(z) = 1/A(z/\gamma)$. From (10), the amplitude vector $\underline{x}$ is given here by:

$$\underline{x}_{opt} = Q^T R \underline{e} \tag{21}$$

Thus, the k elements of $\underline{x}_{opt}$ are simply given by the k largest elements in y of (20). Choosing $0 < \gamma < 1$ results in peak picking from the residual signal colored by $1/A(z/\gamma)$.

We name this scheme Generalized Maximum Residual Magnitude (GMRM) since it is a generalization of simple residual coders proposed in the literature - MRM [5], and TOR (Thinned-Out Residual) [7]. In these coders $\gamma = 0$ is used, and indeed the largest pulses (in absolute value) are picked directly from the residual signal - e_n.

4. An Optimal Vector Set

To suggest an optimal vector set for the above residual expansion scheme, which is independent of the residual signal itself, we modify our previous notation and apply a statistical approach which allows us to use the Karhunen-Loeve (KL) expansion [11] of the speech signal.

Since we are interested in the weighted mean square error (WMSE) it is reasonable to derive the residual-expansion optimal vector set from the KL basis vectors of the weighted speech signal - $\underline{s}_w$, where:

$$S_w(z) \triangleq S(z)W(z) = S(z)\,A(z)/A(z/\gamma) \tag{22}$$

Let $\hat{\underline{s}}_w$ be the approximated weighted speech vector given by:

$$\hat{\underline{s}}_w = \sum_{i=1}^{k} c_i \underline{b}_i \tag{23}$$

where c_i designates the KL expansion coefficient associated with the KL basis vector $\underline{b}_i$. Then, the WMSE is given here by,

$$E_w = E\left\{ (\underline{s}_w - \hat{\underline{s}}_w)^T (\underline{s}_w - \hat{\underline{s}}_w) \right\} \tag{24}$$

where E denotes the expectation operator. The classical result is that E_w is minimized if the basis vectors - $\underline{b}_i$ are the eigenvectors of the matrix $R_{s_w s_w} = E\{\underline{s}_w \underline{s}_w^T\}$, which correspond to its k largest eigenvalues. Due to the orthogonality of the eigenvectors, the KL expansion coefficients are given by: $c_i = \underline{b}_i^T \underline{s}_w$. Hence, $\underline{s}_w$ is represented as:

$$\hat{\underline{s}}_w = \sum_{i=1}^{k} \hat{\underline{s}}_w^i = \sum_{i=1}^{k} \underline{b}_i^T \underline{s}_w \underline{b}_i \tag{25}$$

Writing $\hat{\underline{s}}_w^i$ as $\hat{\underline{s}}_w^i = \tilde{H} x_i \underline{v}_i$, where $\underline{v}_i$ is a member of the residual expansion vector set V and x_i is its multiplying coefficient. Thus, from (25), we obtain:

$$x_i \underline{v}_i = \tilde{H}^{-1}(\underline{b}_i^T \underline{s}_w)\underline{b}_i = (\underline{b}_i^T \underline{s}_w)\tilde{H}^{-1}\underline{b}_i \tag{26}$$

Therefore, with this statistical approach to the problem, the optimal vector set V should be composed of the vectors:

$$\underline{v}_i = \tilde{H}^{-1}\underline{b}_i \quad , \qquad i = 1,2,,\ldots,N \tag{27}$$

where $\underline{b}_i$ are the eigenvectors of the matrix $R_{s_w s_w}$. The matrix Q is obtained then from (27) by selecting those k vectors which correspond to the largest eigenvalues of $R_{s_w s_w}$.

Multiplication by $\tilde{H}^{-1}$ in (27) is equivalent to passing the vectors $\underline{b}_i$ through a filter with the transfer function $A(z/\gamma)$ and truncating the output vector to N samples. Thus, another alternative would be to apply the KL expansion to the weighted signal vector $\underline{s}_w$ and transmit the largest k coefficients. At the receiver the reconstructed weighted signal $\underline{s}_w$ would be passed through $A(z/\gamma)$ (because of $\tilde{H}^{-1}$), in cascade with synthesis filter - $1/A(z)$, to obtain the reconstructed speech signal. Two problems arise with this scheme. First, the receiver needs to know $R_{s_w s_w}$ (or Q), and second, calculating eigenvectors and eigenvalues is a highly complex task. The first problem could be bypassed if we assume that $\underline{e}$ - the residual signal - is white. This assumption is reasonable if we add pitch prediction to the analysis scheme, thus, eliminating the pitch structure from the residual signal.
In that case:

$$\begin{aligned} R_{s_w s_w} &= E\{\underline{s}_w \underline{s}_w^T\} = E\{\tilde{H}\underline{e}\underline{e}^T\tilde{H}^T\} = \\ &= \tilde{H} R_{ee} \tilde{H}^T = g\tilde{H}\tilde{H}^T = gR' \end{aligned} \tag{28}$$

where g is a positive gain term given by the variance of $\underline{e}$. $\tilde{H}$ includes here the effect of the pitch loop and is known at the receiver since $a_i \quad i = 1, \ldots, p$ - the LPC filter coefficients and the pitch loop parameters are transmitted. Therefore, *a modified vector set* is suggested (for performing the KL expansion of $\underline{s}_w$), which is composed of the eigenvectors of the matrix $\tilde{R}$. The receiver then passes the combination of the selected vectors (i.e. $\underline{s}_w$) through $1/W(z)$ in order to obtain the reconstructed speech.

We still have the problem of solving for the eigenvectors and eigenvalues of R' which is usually unacceptably complex. This can be avoided by approximating the KL transform by the DFT or DCT [12], i.e., using a transform which is not data-dependent.

5. Residual Transforming Schemes

Using the DFT or DCT as an approximation to the optimal KL transform is common practice. In our case the transform is applied to the weighted signal $\underline{s}_w$ obtained by passing $\underline{s}$ - the original speech signal - through the weighting filter $W(z)$ given in (3). Following transformation, the k largest absolute valued transform coefficients are selected and their locations and values are coded and transmitted. The receiver then reconstructs the transformed vector from the received coefficients, inserting zeros for the untransmitted ones. This vector is then inverse transformed and the result (which is $\underline{s}_w$) is passed through $1/W(z)$ to produce the reconstructed speech signal. Note that since the transform is not data-dependent, there is no need to perform pitch prediction.

In the above scheme the DCT was found to give better results in comparison with the DFT - as could be expected [13],[14]. However, in simulations we found that using either one of these transforms gives rather unsatisfactory results at the bit rate of 9.6 Kbps and below. The main reason for this is that the reconstructed speech has a "ringing" type of noise added to it. This happens at low bit rates since only a relatively small number of frequency components (tones) are selected, and they dynamically vary from frame to frame.

5.1 Predictive Transform Coding (PTC)

To reduce the ringing effect we have attempted to use a scheme in which the bits available for coding a given frame are distributed among more frequency elements. This is done by using a bit allocation strategy as commonly used in Adaptive Transform Coders (ATC) [15]. Since we code the transformed prediction residual we name this scheme - "Predictive Transform Coding" - PTC.

The above PTC scheme largely reduced the ringing effect and provided better then communication quality reconstructed speech at 9.6 Kbps, and very high quality speech at 16 Kbps. Furthermore, a reduction in complexity is possible, with almost no effect on quality, by determining the bit allocation on the basis of the original speech envelope (instead of the weighted one) and performing the transform on the residual signal $\underline{e}$ instead of $\underline{s}_w$.

Some recently proposed residual coders [9], [10], also transform the LPC residual signal (unlike earlier ATC coders which transformed the input signal directly), and hence fall into the general framework presented here.

6. Summary

The unified mathematical framework presented in this paper was shown to encompass several residual coding schemes which were proposed in recent years. This provides additional insight to the different schemes and also a common basis for their description. It could also be the basis for developing new residual coders by choosing different vector sets and possibly also different vector selection algorithms than presented here. To arrive at an optimal coding scheme, a statistical approach is applied and leads to the more practical Predictive Transform Coder (PTC) in which the residual signal is transformed (e.g., using the DCT) and a bit allocation scheme, similar to that which is used in Adaptive Transform Coders (ATC), is applied for coding the transform coefficients.

References

[1] B.S. Atal, "Predictive Coding of Speech at Low Bit Rates", IEEE Trans. on Comm., vol. COM-30, pp. 600-614, April 1982.

[2] B.S. Atal and M.R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria", IEEE Trans. on ASSP, vol. ASSP-27, No. 3, June 1979.

[3] C.K. Un and D.T. Magill, "The Residual Excited Linear Prediction Vocoder with Transmission Rates Below 9.6 Kbps", IEEE Trans. Comm., vol. COM-23, pp. 1466-1474, Dec. 1975.

[4] B.S. Atal and J.R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates", Proc. Int. Conf. on ASSP, Paris, France, 1982, pp. 614-617.

[5] S.T. Alexander, "A Simple Noniterative Speech Excitation Algorithm Using the LPC Residual", IEEE Trans. on ASSP, vol. ASSP-33, No. 2, April 1985.

[6] E.F. Deprettere and P. Kroon, "Regular Excitation Reduction for Effective and Efficient LP-Coding of Speech", Proc. Int. Cont. ASSP, 1985, pp. 965-968.

[7] A. Ichikawa, S. Takeda and Y. Asakawa, "A Speech Coding Method Using Thinned-Out Residual, Proc. Int. Conf. ASSP, 1985, pp. 961-964.

[8] M.R. Schroeder and B.S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech At Very Low Bit Rates", IEEE Proc. Int. Conf. on ASSP, 1985, pp. 937-940.

[9] T. Moriya, M. Honda, "Speech Coder Using Phase Equalization and Vector Quantization", IEEE Proc. Int. Conf. on ASSP, Tokyo 1986, pp. 1701-1704.

[10] T. Moriya, M. Honda, "Transform Coding of Speech With Weighted Vector Quantization", Proc. Int. Conf. on ASSP, 1987, pp. 1629-1632.

[11] P.A. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach, Prentice Hall, NJ. 1982, (ch.9).

[12] N. Ahmed and R. Rao, "Orthogonal Transforms for Digital Signal Processing", New-York: Springer-Verlag 1975.

[13] Y. Yemini and J. Pearl, "Asymptotic Properties of Discrete Unitary Transforms", IEEE Trans. on PAMI, vol. PAMI-1, No. 4, October 1979.

[14] M. Hamidi and J. Pearl, "Comparison of the Cosine and Fourier Transforms of Markov-1 Signals", IEEE Trans. on ASSP October 1976, pp. 428-429.

[15] R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals", IEEE Trans. ASSP, vol: ASSP-25, pp. 299-309, Aug. 1977.